시뮬레이션 입력데이타 선정 기법

서울대학교 산업공학과

박 진우 교수

Ⅲ. 시뮬레이션 입력확률 분포 선정기법

Ⅲ.1. 분포함수

- 통계이론의 대부분은 분포의 모수(parameter)로 특징지워지는 분포함수에 관한 것이다.
  즉 분포함수의 수학적 형태는 알려져 있고 단지 모수만 未知이므로 그들을 추정해야 한다.

a. 연속 확률 변수

분포함수	확률밀도함수	확률변수 x의 정의역	모수	E[X]	Var[X]
평등분포 Uniform (Rectangular)	$f(x)=\frac{1}{b-a}$	(a,b)	a,b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
삼각분포 (Triangular)	$f(x)=\begin{cases}\frac{2(x-a)}{(b-a)(c-a)} & ,a\le x\le c\\ \frac{2(b-x)}{(b-a)(b-c)} & ,c<x\le b\\ 0 & ,\text{otherwise}\end{cases}$	[a,b]	a,b,c	$\frac{a+b+c}{3}$	$\frac{a^2+b^2+c^2-ab-bc-ca}{8}$
지수분포 Exponential (1,1/λ)	$f(x)=\lambda e^{-\lambda x}$	(0,∞)	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
감마분포 (Gamma (α,β))	$f(x)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}}$	(0,∞)	α,β	αβ	$\alpha\beta^2$
카이자승분포 (Chi Square ($\frac{\gamma}{2}$,2))	$f(x)=\frac{1}{\Gamma(\frac{\gamma}{2})2^{\frac{\gamma}{2}}}x^{\frac{\gamma}{2}-1}e^{-\frac{x}{2}}$	(0,∞)	γ	γ	2γ
얼랑분포 (n-Erlang(β))	Gamma(n,β)				
정규분포 (Normal)	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	(-∞,∞)	μ,σ	μ	σ^2
t분포	$f(x)=\frac{\Gamma(\frac{\gamma+1}{2})}{\sqrt{\pi\gamma}\,\Gamma(\frac{\gamma}{2})}(1+\frac{x^2}{\gamma})^{-\frac{\gamma+1}{2}}$	(-∞,∞)	γ	0	$\frac{\gamma}{\gamma-2}$ 단(γ>2)
F분포	$\frac{\chi^2_{\gamma_1}/\gamma_1}{\chi^2_{\gamma_2}/\gamma_2}=f(x)=\frac{\Gamma(\frac{\gamma_1+\gamma_2}{2})\gamma_1^{\frac{\gamma_1}{2}}\gamma_2^{\frac{\gamma_2}{2}}x^{\frac{\gamma_1}{2}-1}}{\Gamma(\frac{\gamma_1}{2})\Gamma(\frac{\gamma_2}{2})(\gamma_2+\gamma_1 x)^{\frac{(\gamma_1+\gamma_2)}{2}}}$	(0,∞)	γ1,γ2	$\frac{\gamma_2}{\gamma_2-2}$ 단(γ1>1)	$\frac{\gamma_2^2(2\gamma_2+2\gamma_1-1)}{\gamma_1(\gamma_2-2)^2(\gamma_2-\gamma_1)^2}$ 단(γ2>2)
베타(Beta)분포	$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	(0,1)	α,β	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

- 일반적으로 알려진 감마(Gamma)함수의 형태는 다음과 같다.

$\Gamma(p)=\int_0^{\infty}x^{p-1}e^{-x}dx$ defined for p > 0

$\Gamma(p)=(p-1)\Gamma(p-1)$, $\Gamma(1)=1$ 이므로 $\Gamma(p)=(p-1)!$ 또한 $\Gamma(\frac{1}{2})=\sqrt{\pi}$

- 평등분포는 베타(Beta)분포의 특수형태 : U(1,0) = Beta(1,1)

b. 이산 확률 변수

분포함수	확률변수	확률변수 x의 정의역	모수	E[X]	Var[X]
베르누이 (Bernoulli)	$p(x)=p^{1}q^{1-x}$ 단, q=1-p	0,1	p	p	pq
이항분포 (Binomial)	$p(x)=\binom{n}{x}p^{x}q^{n-x}$	0,1,...,n	n,p	np	npq
초기하분포 (Hyper -Geometric)	$p(x)=\binom{A}{x}\binom{N-A}{n-x}/\binom{N}{n}$	0,1,...,n	N,A,n	$n(\frac{A}{N})$	$n(\frac{A}{n})(\frac{N-A}{N})(\frac{N-n}{N-1})$
음이항분포 (Negative Binomial (Pascal))	$p(x)=\binom{x-1}{r-1}p^{r}q^{x-r}$	r,r+1,...	r,p	$\frac{rq}{p}$	$\frac{rq}{p^{2}}$
기하분포 (Geometric)	$p(x)=pq^{x-1}$	1,2,3,...	p	$\frac{q}{p}$	$\frac{q}{p^{2}}$
포아손 (Poisson)	$p(x)=\frac{\lambda^{x}e^{-\lambda}}{x!}$	0,1,...	λ	λ	λ

- 기하분포는 음이항분포에서 r=1의 경우, 즉 첫번째 성공까지의 시행의 횟수이다.

- 추계학에서의 Counting Process에서 다음의 3가지 가정을 하여주고 미분방정식으로 풀어주면 포아손 분포(Poisson Distribution)가 나온다.

(가정) 1. 작은 구간내에서 사건이 발생할 확률은 거리에 비례
2. 한 구간내에서 2개의 사건이 발생할 확률 극소
3. 구간이 겹치지 않으면 독립

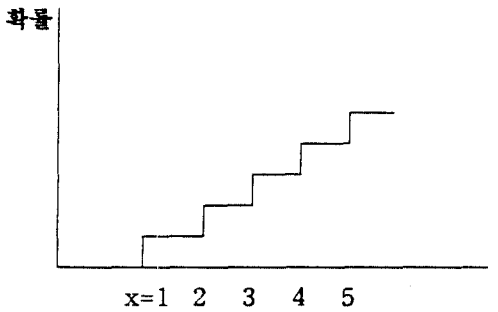
$$p(x)=\binom{n}{x}p^{x}q^{n-x} \quad \frac{np=\lambda}{n\to\infty} \rightarrow \quad \frac{e^{-\lambda}\lambda^{x}}{x!}$$

포아손분포는 러시아에서 말에 채어 죽은 사람의 비율을 조사하다 만들어 졌다.

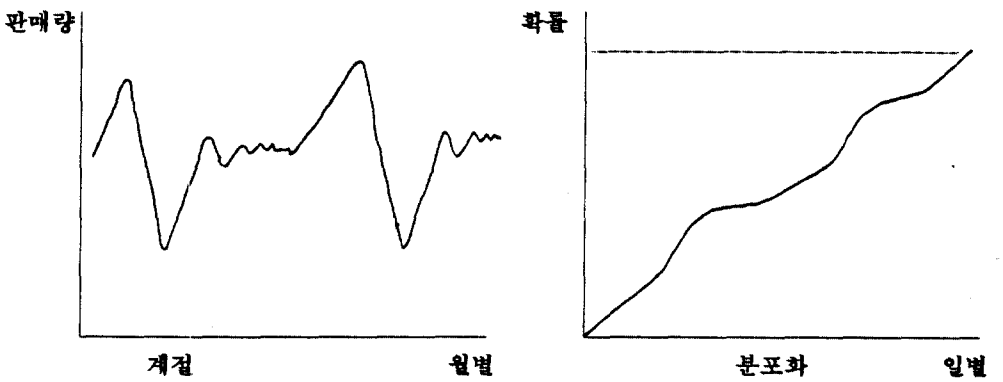
c. 경험적 분포

- 이른바 비모수분포의 예이다.

예1) 신문팔이 소년의 문제 : 계단함수이면서 비감소함수이다.



예2) 년간 판매량 또는 년간 누적 판매량도 하나의 분포함수가 될 수 있다.



Ⅲ. 2. 분포의 선택 (어떤 확률적 현상 보았을때 그 같은 현상을 나타나게 하여준 숨어주는 분포의 추정방법)

- 일반적으로 경험적분포보다는 밀도함수가 사용하기에 편하다. 왜냐하면 수개의 모수만으로 분포가 완벽하게 정의되기 때문에 메모리를 절약할 수 있는 등의 장점이 있다.

- 연속 또는 이산 분포여부는 확률변수 특성에 따라 쉽게 구분이 가능하다.

a. 점 추정을 이용하는 방법

- 첫번째, 두번째 모멘트는 확률적으로 발생하는 관측치를 이용하여 쉽게 계산하여 쓸수 있다.

첫번째 모멘트 = $\sum_{i=1}^{n} \frac{X_i}{n}$, 두번째 모멘트 = $\sum_{i=1}^{n} \frac{(X_i - \bar{x})}{n-1}$

1) 연속형 확률변수

(정의) 분산 계수 = 표준편차 / 평균

(예) Exp(λ) $f(x) = \lambda e^{-\lambda x}$

$E[X] = \frac{1}{\lambda}$, $VAR[X] = \frac{1}{\lambda^2}$, $Cov = 1$

2) 이산형 확률변수

이경우는 분산계수보다 분산/평균이 더 유용하다.

예) 이항분포 $Bin(n,p)$: 분산 = npq , 평균 = np

따라서 분산/평균 = $\frac{npq}{pq}$ $\in (0,1)$

예) 기하분포 $Geo(p)$: 분산 = q/p^2 , 분산 = q/p

따라서 분산/평균 = $1/p$ $\in (1,\infty)$

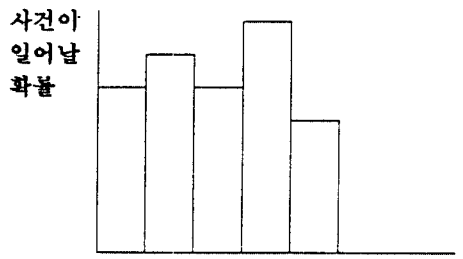
예) 음이항분포 $Neg(n,p)$: 분산 = nq/p^2 ,평균 = nq/p

따라서 분산/평균 = $1/p$ $\in (1,\infty)$

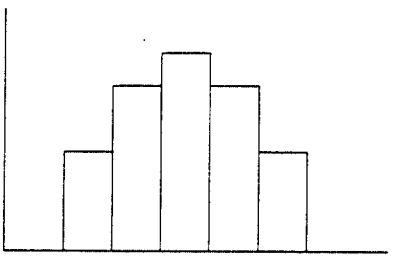
예) 포아손 $Poisson(\lambda)$: 분산 = λ , 평균 = λ

$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$

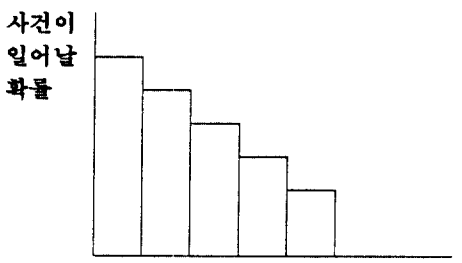
b. 히스토그램(Histogram)



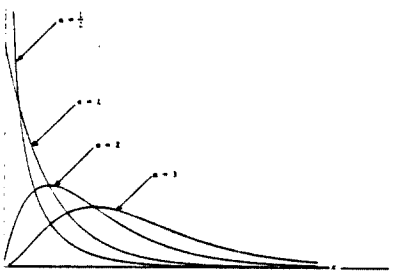
평등분포



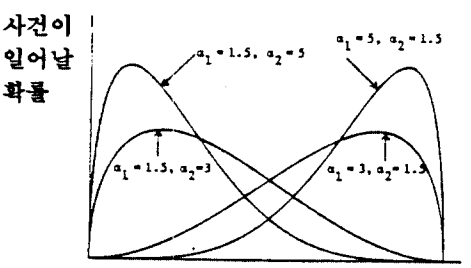
정규분포



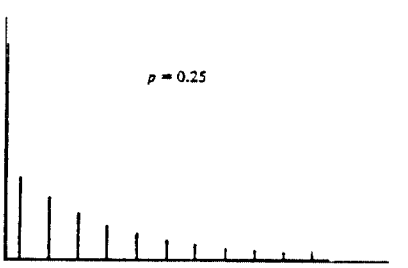
지수분포



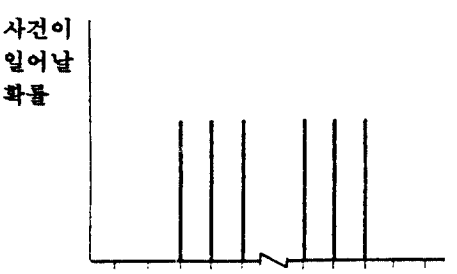
감마분포



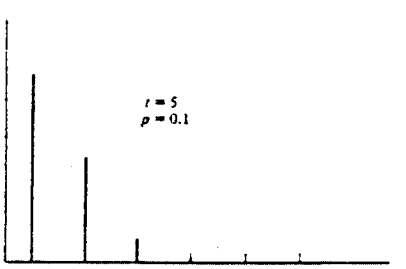
베타분포



기하분포



이산 평등 분포



이항분포(음이항분포,포아손분포)

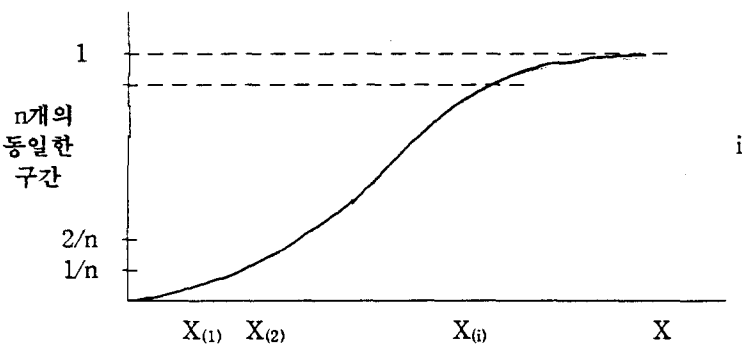
히스토그램을 통해 모수의 특정까지 추정이 가능

모수 ┌ 위치 ⇒ x, x-r, x+r
　　 ├ 규모 ⇒ 퍼짐의 정도
　　 └ 형태 ⇒ 위의 두 모수보다 더 기본적인 모수로 지수와 정규분포등과 같이 형태가 완전히 다른 경우, 분포는 형태를 모수로 가지지는 않는다.

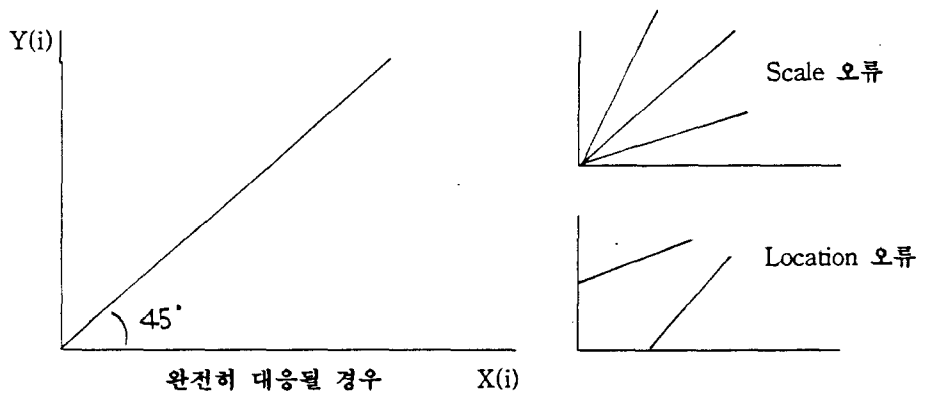
c. 확률 도표(Probability Plots)

(기본 아이디어)
거의 모든 분포는 S형태를 한다. ---> 선형화시켜서 비교.

(알고리즘)
1. n개의 샘플이 주어졌을 때, 증가순으로 재배열하여 $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$을 순서통계량이라고 한다.
2. 어떤 그럼직한 분포(모수가 결정된)에서 정의역을 n개의 동일구간으로 분할하고 i = 1, 2, ... , n 에서의 X(i)의 특정값을 발견할 수 있다.
(Note) 분포가 완전히 결정되어야 함.



3. Y(i)에 대하여 X(i)에 대한 그래프를 도시함.



Ⅲ. 3. 적합도 검정 (정량적 방법)

- 여기서 χ^2 검정과 콜모고로프-스미르노프(Kolmogorov-Smirov)검정을 배우게 되는데 이들은 이른바 비모수 통계적 추론의 예이다. (즉 여기서 우리가 취급하는 것은 분포의 모수집단을 가정하여 두고, 의심가는 모수를 결정하는 일반적인 검정이 아님)

χ^2 검정	Kolmogorov-Smirov 검정
연속 및 이산 확률변수 공통	연속 확률변수만

a. χ^2-검정

- 구간의 관측된 빈도 수와 기대되는 빈도 수를 비교한다.

만일 모집단이 K개의 다른 타입으로 구성되어 있다고 하자.

즉, P_i = 타입 i를 선택할 확률. i = 1,...,K 로 두면

$$P_i \geq 0 \quad i=1,...,k \quad \sum_{i=1}^{k} P_i = 1 \text{ 가 성립한다.}$$

이때 다음과 같이 가설을 정하자.

H_0 : $P_i = P_i^o \quad i=1,...,k$

H_1 : 적어도 하나 이상의 i값에 대하여 $P_i \neq P_i^0$

가설을 검정하기 위하여, 샘플을 취하여 보자(관측을 한다).

N_i = 타입 i인 관측의 횟수

$\sum_{i=1}^{k} N_i = n$ (총 n개의 샘플)로 정의하자.

문제 자체는 비모수 문제가 아니다. 그러나, 많은 비모수 통계가 이런 유형의 문제에 적용될 수 있다. 만일 H_0이 사실이라면, 타입 i인 관측치는 $n P_i^0$ 와 같을 것이며, 따라서, 그 차인 $(N_i - n P_i^0)$는 작을 것이다.

1900년경 유명한 통계학자 K. Pearson은 다음과 같은 통계량을 제안했다.

$$Q = \sum_{i=1}^{k} \frac{(N_i - n P_i^o)^2}{n P_i^2}$$

또한 그는 만일 가설 H_0가 사실이라면 Q는 n이 무한대로 수렴함에 따라 Q는 $\chi^2(k-1)$의 분포를 따른다는 사실을 증명하였다.

(Test의 절차)

1. 가설 H_0를 세운다.
   즉, $F(x) = F_0(x)$: 단 여기서 $F_0(x)$는 현상을 나타내었으리라고 생각되는 그럴듯한 분포.

2. $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}$ 를 구한다. (k개의 구간으로 나눈다.)

3. 만일 $\chi^2 \le \chi^2_{\alpha;\nu}$ 이면 가설 H_0를 채택한다.

   a :유의 수준 (Level of Significance)
   $\nu = k-1$: 자유도 (k-1)

-만약 선택된 분포 $F_0(x)$가 미지의 모수 $\Theta=(\theta_1, \cdots, \theta_s)$, s<k-1,을 포함할 경우의 χ^2의 분포는?
Ω를 Θ의 모든 가능한 값을 포함하는 S-차원의 모수공간이라 하자.
모든 $\Theta \in \Omega$라 가정하면, 함수 $\Pi_i(\Theta)$는 확률 P_i 의 가능한 값을 나타낸다. 즉,

$\prod_i(\theta) \ge 0$, $\sum_{i=1}^{k} \prod_i(\theta) = 1$.

H_0 : $P_i = \Pi_i(\Theta)$, i=1,…,k 인 $\Theta \in \Omega$가 존재한다.
H_1 : H_0가 거짓이다.

의 가설에 대한 검정이 필요하다. 그런데, 1924년 유명한 통계 학자 Fisher는 다음의 정리를 발표하였다.

[정리] (1924, R.A. Fisher)

만약 $\hat{\theta}$이 최우추정치(Maximum Likelihood Estimator)이고 $\hat{\theta}$이 모종의 균일화조건(Regularity Condition)을 만족시키면

$Q = \sum_{i=1}^{k} \frac{[N_i - n\Pi i(\hat{\theta})]^2}{n\Pi i(\hat{\theta})}$ 는 n이 무한대로 증가할때 $\chi^2(k-1-s)$의 분포를 따른다.

(예) 정규분포 경우의 균일화 조건은 예를 들면
정규 정규분포의 우도함수 $L(\mu,\sigma^2) = [\Pi i(\mu,\sigma^{2})]^{N_1} \cdots [\Pi k(\mu,\sigma^{2})]^{N_k}$에서 $X_1,\ldots,X_n$. 로부터 얻어진 최우수 추정치를 사용하지 않는다는 것이다.

우리는 $\hat{\mu} = \overline{X}$, $\hat{\sigma}^2 = \sum \frac{(X_i - \overline{X})^2}{n}$ 임을 최우추정치로부터 구할 수 있다. 그러나, 균일화조건을 만족시키기위하여는 이같은 $\hat{\mu}$나 $\hat{\sigma}$를 사용하여서는 않된다는 것이다. 즉, k개의 구간으로 나누었을때 각 구간에 속하는 사건수를 N_i라고 하자. 만약, μ, σ^2이 未知라면 $\theta = (\mu,\sigma^2)$이다.

그러면, $\pi_i(\mu,\sigma) = \int_{a_i}^{b_i} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$ 가 된다.

이경우 X_i가 아닌 단지 N_i만을 미지로한 우도 함수(Likelihood Function) μ, σ^2를 구하라

[정리] (1954, H.Chernoff & E.L.Lehmann)

만일 최우추정치 $\hat{\mu}$ 와 $\hat{\sigma}^2$ 이 사용되고, H_0가 사실이라면
Q는 $\chi^2(k-1)$ 과 $\chi^2(k-s-1)$사이에 놓이는 분포함수에 수렴한다.
그래서 보다 보수적인(안전한) χ^2검정에서는 만일 $\chi^2 \le \chi^2_{\alpha,k-1-s} \le \chi^2_{\alpha,k-1}$ 이 만족되면 가설 H_0를 채택하고 $Q > \chi^2_{\alpha,k-1}$ 이면 기각하는 절차를 권장하고 있다.

(χ^2-Test 의 문제점)

① 어떻게 구간의 수 k와 구간의 크기를 정할것인가?

② 같은 표본으로부터도 相異한 결과가 가능하다.
(문제점에 대한 경험적 해결 방법)

① 각 구간당 최소 5개의 관측치가 포함되도록 한다.
(1.5 개도 만족할만하다.)

nP_i° 가 아주 작지 않은경우에는, χ^2 는 Q에 좋은 근사치가 된다.
부연하자면 대략적으로 $nP_i^\circ \ge 5$인 경우에는 근사가 매우 만족할만 하고, $nP_i^\circ \ge 1.5$ 인 경우에도 역시 만족할만 하다.

② E_i의 값이 가급적 균등하도록 구간을 나누어준다.

③ 구간의 수 k는 $k \le 30$ (or 40)을 만족시키도록 한다.

b. 콜모고로프--스미르노프 검정

$X_1,...,X_n$ 을 어떤 연속 확률 분포로부터의 표본들이라 하자. 이때 $x_1,...,x_n$ 을 $X_1,...,X_n$의 관측치라고 두면 각각의 X_i들은 연속분포를 따르기 때문에,

$\Pr(X_i = X_j) = 0$ ==> 즉 각 x_i들은 서로 다르다고 가정할 수 있다.

이제 이들 표본으로부터 다음과 같은 표본분포함수 $F_n(x)$을 정의한다.

$F_n(x)$ = 표본들 중에서 x보다 작거나 같은 관측치의 비율

즉, $F_n(x) = \dfrac{Number\ of\ x_i's \le x}{n}$

또한 F(x)를 확률표본이 추출된 분포라 하자.

그러면 대수의 법칙에 의하여,

$$p\lim_{n\to\infty} F_n(x) = F(x) \qquad -\infty < x < \infty$$

(참조) 확률적 수렴

$$\lim_{n\to\infty} \Pr\{|F_n(x) - F(x)| < \varepsilon\} = 1, \quad for\ any\ given\ \varepsilon > 0$$

[글리벤코-코텔리 정리] 평등수렴에 관한 정리

$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|$,이라 두면 $p\lim_{n\to\infty} D_n = 0$가 성립한다.

(콜모고로프-스미르노프 검정 절차)

1. 귀무가설 H_0 : $F(x)=F_0(x)$ 을 세운다.
2. 확률표본으로부터 표본확률분포를 만든다.
3. 검정할 통계량을 구한다.

즉, $D_n = \underset{-\infty<x<\infty}{Sup} |F_n(x)-F_0(x)|$

$X_1,...,X_n$의 값들이 관측되기 전까지는, D_n은 확률변수이고, 만약 H_0가 참이면 D_n은 $F_0(x)$와 독립인 분포를 따른다.

D_n의 분포함수는 n과 유의수준 α에 관한 표(밑에 언급된 $d_{n,\alpha}$는 이같은 표로서 만들어 줄 수 있다.)로 나타낼 수 있다.

4. $D_n(x) > d_{n;\alpha}$ 이면, H_0를 기각한다.

대략적으로 $d_{n,10\%} = \frac{1.22}{\sqrt{n}}$

$d_{n,5\%} = \frac{1.36}{\sqrt{n}}$

$d_{n,1\%} = \frac{1.63}{\sqrt{n}}$ 이다.

위의 검정절차중 3및 4는 1930년 발표된 A.N.Kolmogorov 와 N.V.Smirnov의 연구결과에 그 근거를 두고 있기때문에 이 검정절차를 콜로고로프-스미르노프 검정이라 칭한다.

[정리] (1930 A.N.Kolmogorov & N.V.Smirnov)

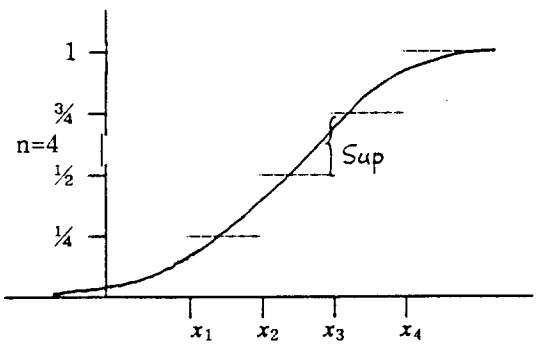
H_0가 참이면, 어느 주어진 임의의 $t>0$ 에 대해서,

$$\lim_{n\to\infty} \Pr(n^{1/2}D_n \le t) = 1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}$$

$n^{1/2}D_n > Cn^{1/2}$일때 H_0를 기각하는 테스트과정을 콜모고로프-스미르노프 검정이라 한다.

(콜모고로프-스미로노프 검정에 대한 토의)

① $F_n(x)$: 오른쪽 연속 계단함수이기 때문에 아래 그림과 같이 D_3을 $F_n(X_3)-F_0(X_3)$로 할 것인지 $F_n(X_2)-F_0(X_3)$으로 할 것인지의 문제가 생긴다.



실제응용에서는 모든 i에 대해서 $|F_0(x_i)-F_n(x_i)|$ 와 $|F_0(x_i)-F_n(x_{i-1})|$를 찾아주고 이들 2n개의 편차중 가장 큰 것을 D_n으로 둔다.

② $F_0(x)$는 연속이고, 완전히 정의되어야만 K-S검정을 적용할 수 있다. 이때 시뮬레이션에 의해 미지 모수 경우의 기각치가 구해진다. (정규,지수,와이불 분포의 경우)

(2개의 표본을 위한 K-S 검정)

X_i (i=1,...,m) 를 미지의 연속 분포함수 F(x)에서의 확률표본이라 하고,
Y_j (j=1,...,n) 을 미지의 연속 분포함수 G(y)에서의 확률표본이라 하자.

이 경우의 검증 절차는 다음과 같다.

1. 귀무가설 : $F(x) = G(x) \quad -\infty < x < \infty$
   대립가설 : $F(x) \neq G(x) \quad -\infty < x < \infty$ 을 세운다.
2. $F_n(x)$, $G_n(x)$의 표본 분포함수를 만든다.
3. 검정통계량

$$D_{mn} = \sup_{-\infty < x < \infty} |F_n(x) - G_n(x)|$$ 을 구한다.

4. $D_n(x) > d_{n;\alpha}$ 이면 귀무가설을 기각한다.

   단, 여기서 $d_{n;10\%} = \left(\frac{mn}{m+n}\right)^{-1/2}$, 1.22

Ⅲ. 4. 기타 주제

a. 만약 분포를 검정할만한 표본이 없을 경우

최소한 가능한 구간만이라도 추정 Uniform(a,b) ┐
구간 + 최빈값 (정점)만 알 경우 Triangular ├ 를 시도한다.
구간 + 최빈값 + 평균만 알 경우 Beta ┘

b. 고객의 도착 분포

Poisson Process : $E[N(S)] = \lambda S$

여기서 N(S)는 구간 (0,S)동안의 도착횟수

λ 는 도착률(단위 구간동안의 도착수의 기대값)

동태적(Non-Stationary) Poisson Process : 시간 의존적인 도착율 $\lambda(t)$

그룹의 도착(Batch arrival) : 복합 포아슨 분포로 묘사한다.